Distillation and Amalgamation

Dr. Ali Arsanjani

Machine Learning Week 8 (c)
DeepContext 2016-2019

10/13/20

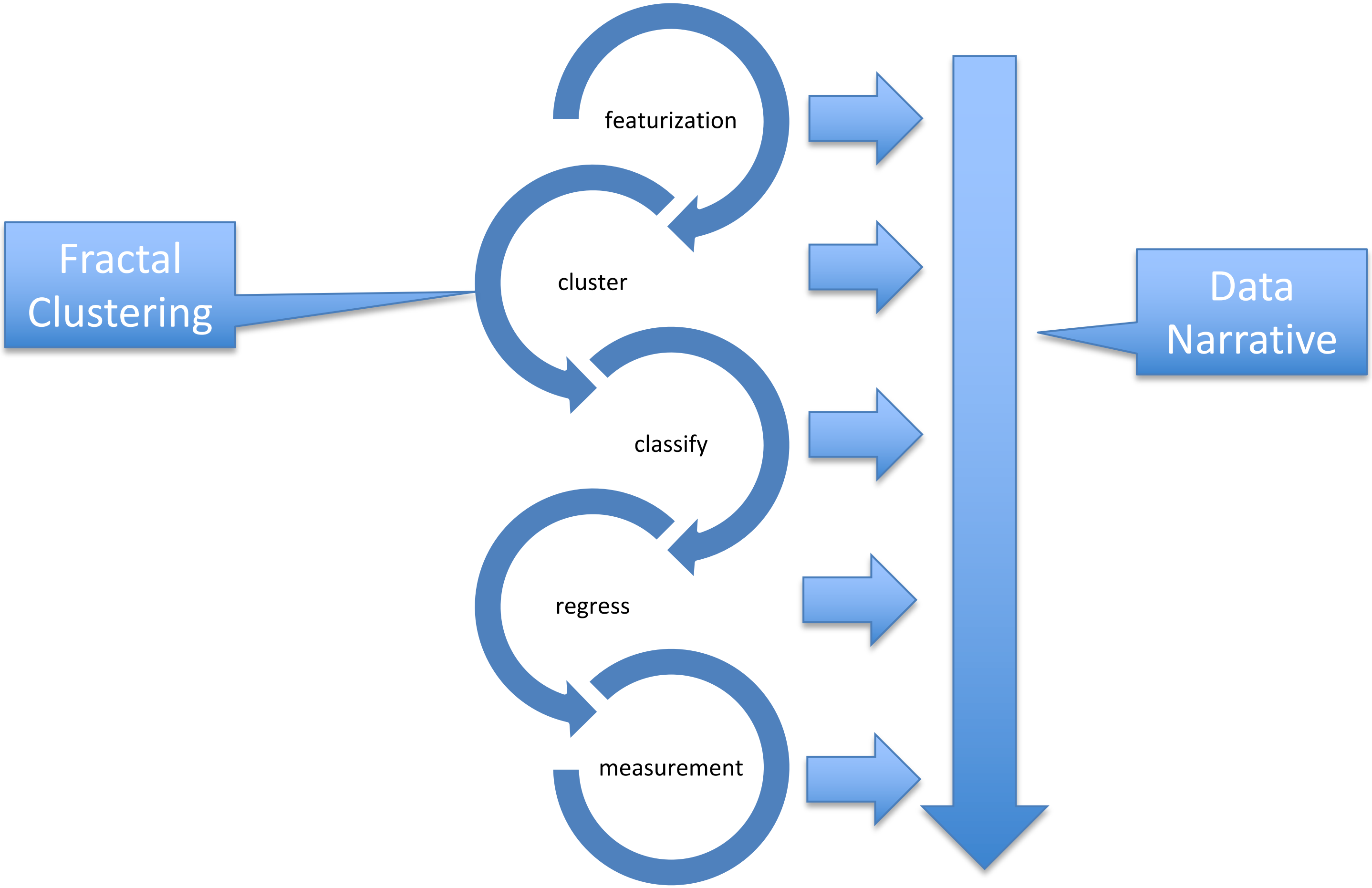
ML

- Clustering, Classification, regression
- Recommendations
- Forecasting (time-series)
  - ARIMA, unique algos

Experiments → Questions

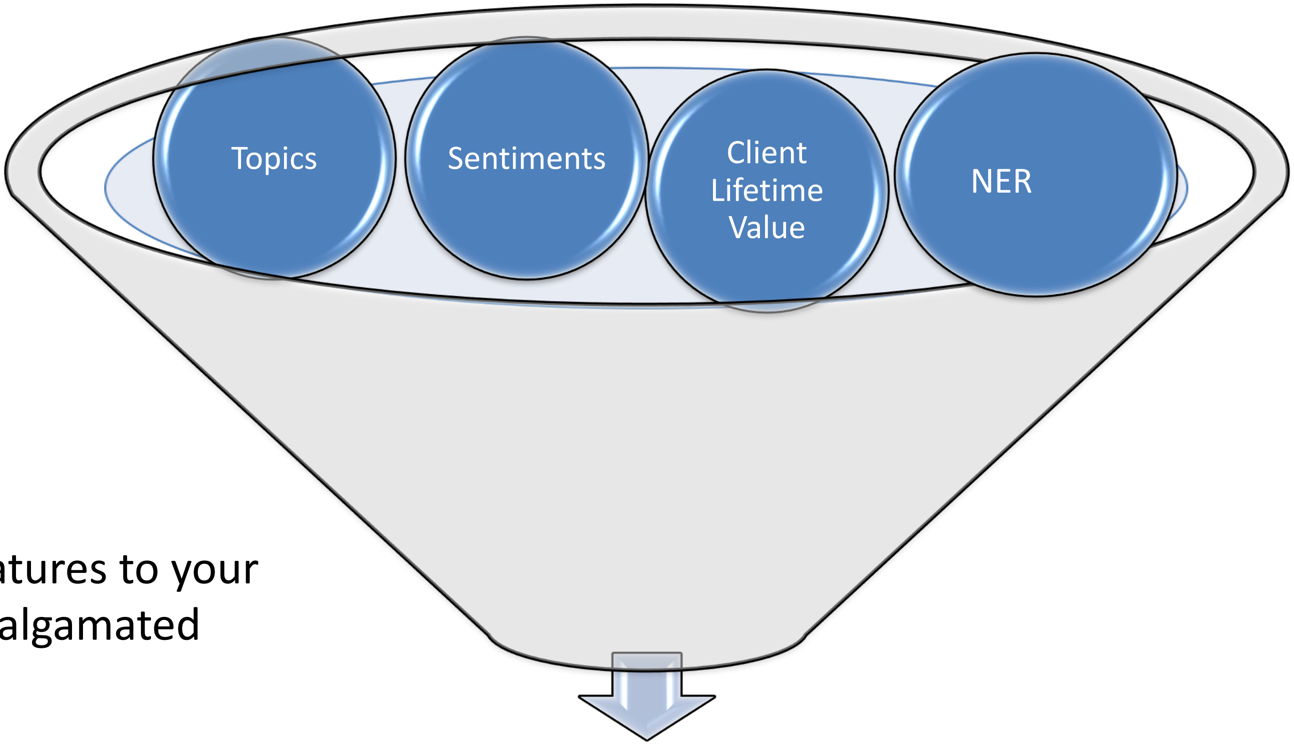
- Number of questions your project answers
- Data Narrative → Business Goals, Objs
- SWE : Prototypes: MVP , Sprints1

The ML Life-cycle is a Journey of Increased Refinement



Distillations

NLP



Add new features to your existing, amalgamated dataset

Distilled Contextual Embedding

Add more accuracy, precision, recall, f1 , RMSE, CM

Distillations

10 insights

Distillations

1. [Entity] Customer Identity
   - (basic, lookup: e.g., caller id → pull up record in CRM)
   - Identity of the group
   - Customer hyper segmentation and hyper personalization → fractal clustering
   - Characteristics of the golden cluster → propagate to entire data set
2. Entity Resolution
   1. Amalgamation:
      1. Embeddings, Cosine Distance (words that have meanings close to one another, or occur in similar **contexts**)
      2. Literals, Euclidean distance (e.g., geolocation to find common communities, areas, etc)
   2. E.g., Senzing.com
   3. MDM master data management → entity resolution
      1. Do these rows in diff datasets refer to the same customer?
3. Customer Lifetime Value (complex), Customer Rank (simple)
   - E.g, simple : how much have they purchased to date?
   - How much do we anticipate (regression) they will purchase in the future based on prior purchases (time-series)? What is their propensity to buy? To convert lead to a customer ? What is their propensity for an upsell/cross-sell?
4. Sentiment Analysis
   1. Sentiment a la *Vader*
   2. Tone Analysis
   3. Personality Insights
   4. BERT for Sentiment Analysis;

Distillations

1. [Entity] Customer Identity
2. Entity Resolution
3. Customer Lifetime Value (complex), Customer Rank (simple)
4. Sentiment Analysis
5. Topics
6. Requests/Intent
7. Time lines
8. Locations
9. Entities and Relationships Extraction (Named Entity Recognition (NER) Extraction)
10. Dictionary, Ontology

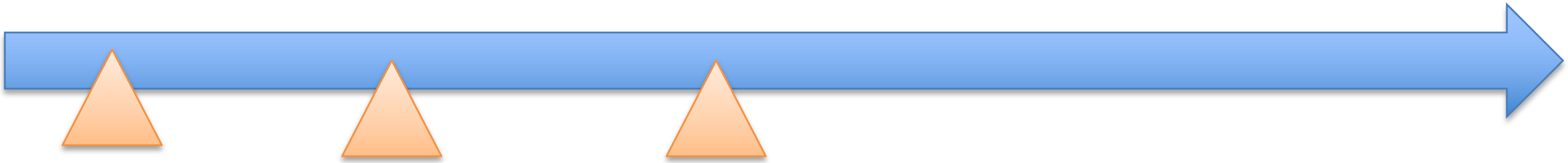
Distillations

5. Topics
   - LDA, LSA, variations
     1. Attention based LSTM with n-grams
     2. BERT
6. Requests/Intent
   - Use topics to unearth actual requests customers are making, what are they asking when they call, text, email?
7. Time lines
   - Construct an event timeline of the customer behavior, patient case, insurance claim, stock symbol relative to market, realestate prices, etc.
8. Locations
   - Geolocation, community, lat-long, zipcode, etc.
9. Entities and Relationships Extraction (Named Entity Recognition (NER) Extraction)
   - For knowledge graph construction
   - NER
10. Dictionary, Ontology
    - construct a lexicon, taxonomy, of jargon
    - Ontology
    - Knowledge graph

Distillations

Forecasting
Time series

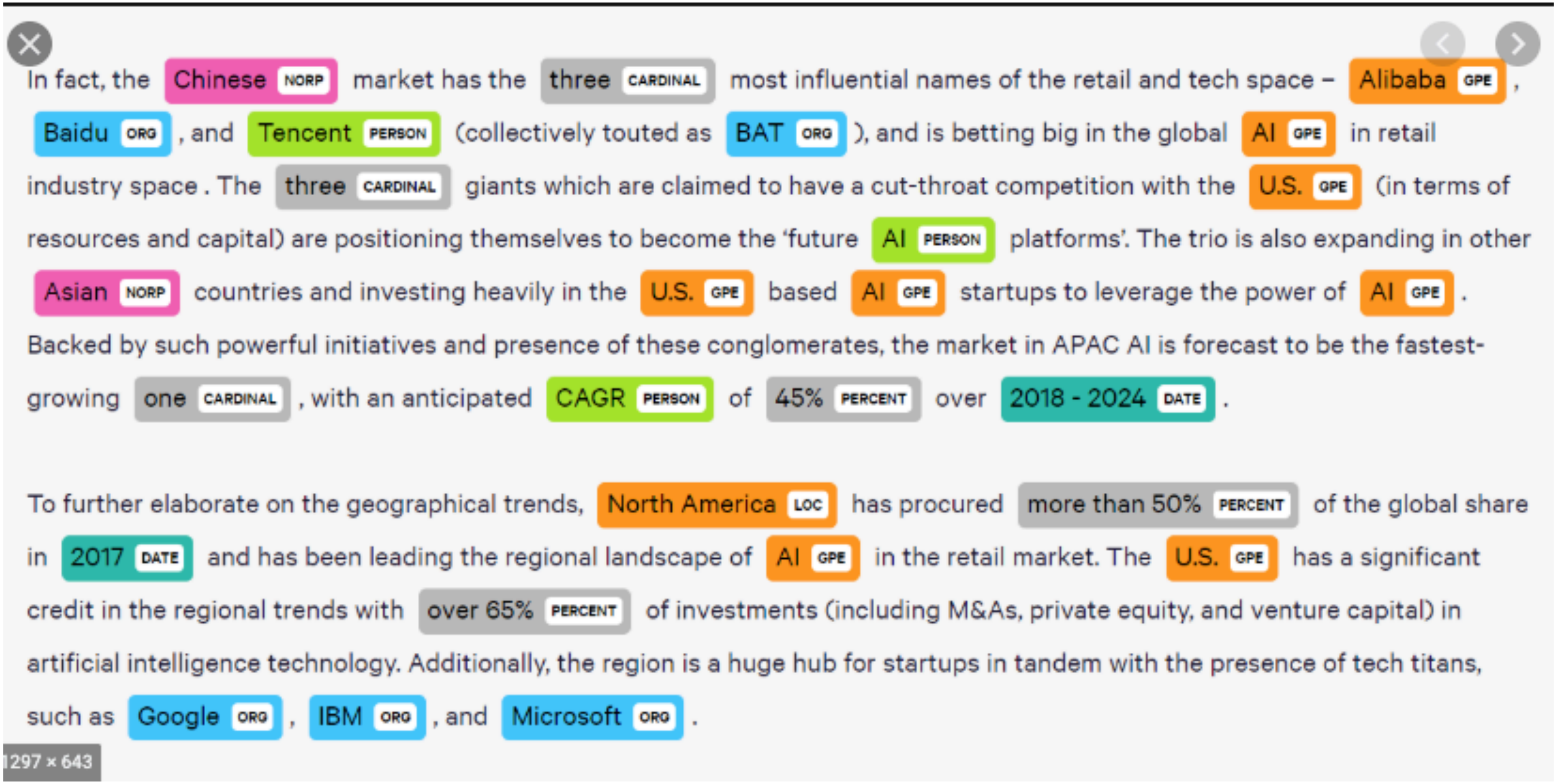
Recomm for next best
action



Each Team: Pick one Distillation to research and provide report and example using the same data set

- Find and scrape data for your 3rd distillation
- Find text related to your dataset
  - Reviews?
  - News? In that industry, about that company?
  - Government published data related to it
  - Disinformation related to it!?
- Can you do this distillation with a given dataset?
  - If yes how?
  - If not, why not and how can we get close to that distillation?
    - Eg Customer Identity
- Write various design options of doing that distillation
- Choose an implementation with code

NER: Tags Known Entities with Meta-data



In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE , Baidu ORG , and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail industry space . The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the 'future AI PERSON platforms'. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE .

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG , IBM ORG , and Microsoft ORG .

Adding new entities and meta data

- That is not known by the NER
- Custom NER
    - Dict : { “<entity word>” : “<label>”}
- Topics are input for your dict

Machine Learning Week 7